Addressing circular definitions via systems of proofs

Addressing circular definitions via systems of proofs

Riccardo Bruni*

Abstract

Definitions are important tools for our attempt to construct an intelligible imagine of reality. Regarded as such, there are interesting epistemological clues about them to consider: What does a legitimate definition look like? Or: Is there a privileged form that definitions should have? How do we find good definitions? Or: How do we know that a given definition is a good one? Traces of the debate on these, and similar other questions are ubiquitous in the history of philosophy. The aim of this note, however, is not to give a historical account on this matter. Rather, on the basis of some very recent work of a proof–theoretical nature, I plan to address those issues in the context of a discussion concerning circular definitions. A definition is circular in the sense of this paper if the very concept that one is defining is used in the condition defining it. Despite their peculiar character, circular definitions are neither rare, nor easy to dispense with: it turns out that they affect in a significant way the ordinary life, as well as the philosophically interesting level of speech. That is the very reason why they have slowly gathered consideration from scholars in recent times, and have become matter of debate. Circular definitions break the traditional schema that can be used to give an account for ordinary definitions, therefore they raise the problem whether they are legitimate or should be avoided. The goal of this paper is to discuss the issue of circular definitions, illustrate the problematic features connected to it, and present recent developments in the logical research on the topic that help providing them with an arguably plausible justification.

Keywords Circular definition, revision theory (of circular concepts), proof theory, *mathesis universalis*

1 Introduction

The act of defining concepts is the very act by means of which things become 'manageable' to reason. Actually, a definition can, or should be viewed as an

*Dipartimento di Lettere e Filosofia, Università degli Studi di Firenze, via Bolognese 52, 50139 Firenze, Italy. This work was supported by the Italian Ministry of Education, University and Research through the PRIN 2017 program "The Manifest Image and the Scientific Image", prot. 2017ZNWW7F_004.

act of understanding itself, since it aims at making explicit our knowledge of something (or to explain the way in which our use of a term for referring to something is to be understood). Definitions are, so to say, milestones along the path toward our making reality intelligible. Regarded as such, they give rise to a number of questions, those closest to the spirit of the present contributions being about (i) whether there are boundaries to our activity as long as the definition of "things" is concerned (that is, whether one should see this activity as being limited, somehow *a–priori*, due to factors depending upon how reality is made, or how reason is made, etc.), (ii) whether there are formal criteria that help establishing the difference between good and bad definitions.

Granted the crucial role that definitions play for our intellectual activity, it does not come as a surprise to discover that references to definitions are ubiquitus, particularly in the philosophical literature. Given the quite modest aim that I am pursuing with the present contribution, to provide the reader with a detailed account for that would be by far off topic. To briefly go through the issue with a historically-oriented look instead, might be useful to set down the basis for my subsequent discussion. In particular, to have a look at the traditional concept of definitions as the latter turns out from classical texts in philosophy could serve the purpose of introducing the peculiar viewpoint I will pursue.

According to knowledgeable scholars in the field, Aristotle's view of definitions, as is often the case, has turned out to be seminal. The view he advances in his *Posterior Analytics* amounts to the idea that good definitions should have a two-fold objective: they should in the first place capture the essence of what is defined; they should provide us, in addition, with an explanation of the causes of it (see [7, pp. 136-138]). In other words, a good definition should permit to exhaustively answer two questions: "*What* is that makes something what it is?", and "*Why* is something what it is?". Since history is not my main concern, I consent myself to freely elaborate a little bit this idea for the sake of argument. The main point of it that is worth stressing in my opinion, is that by speaking of "essence" and "causes" one ends in a philosophically non-innocent viewpoint. Definitions are linguistic acts and they could be regarded as solely fixing the meaning of terms (see the next section). However, languages can be legitimately viewed as tools we use to speak of reality, or its "portions", and that is what essence and causes are about. So, to assume that definitions should reflect features which are proper to an element of the outer reality, be that physical or conceptual in nature, seems to entail a view that departs from a more neutral approach, ontologically speaking, based upon primarily stressing the purely "linguistic character" of definitions.

Aristotle was of course fully conscious about this aspect of the issue. As a matter of fact, he clearly distinguishes between the role of definitions in explaining the meaning of a term, from the role they play in establishing the essence of what the term signifies (see again [7, p. 135, p. 141]). However, it seems that in his view these roles are strictly connected, as they are both tied up with the goal of the scientific investigation in the sense that they provide stages in the process according to which knowledge of reality is attained. The process leading

to the latter, according to Aristotle, features a first stage in which definitions are just taken as linguistic tools that fix the meaning of terms; passes through a stage at which one knows that terms correspond to existing "objects"; and ends in a final stage which is achieved when one reaches consciousness that by the original definition one has successfully captured the essence of the thing it defines (see [7, pp. 137-142]). To exemplify this as it is done in [7] for the sake of self-containedness, this means that a definition like:

Triangles are three–angled plane figures

should be rather presented at the first stage in the three-step process referred to above in the following form:

Triangles, *if they exist*, are three–angled plane figures

The statement certifying that stage two has been reached, would then read as:

Triangles *exist as* three–angled plane figures

and it would lead to a third and final form of the original definition, which reads like the corresponding universal statement:

All triangles are three–angled plane figures

As a consequence of this view, definitions as they are first grasped have a form and a content that is significantly different from the most profound way in which they are understood at last. In particular, it follows that they should not be immediately regarded as universal statements. For, a statement of this latter sort already comprises for Aristotle an assertion of existence, which is not necessarily included in a definition as the latter occurs in the first stage. Therefore, definitions in this schema are *prima facie* different from universal statements, and require a philosophical assessment on their own[1].

As it was said, Aristotle's viewpoint and analysis of definitions was not isolated among Greek philosophers. Rather, it seems that both Plato's analysis, as well as the one proposed by the Stoics, were in agreement with the most significant aspects of it[2]. Then, some questions rise spontaneously: what makes a definition a good one? If we, as definition makers, are supposed to isolate the essence of things, could our activity in this sense be ruled out? Could it be possible to formulate criteria, of either formal or contentual character, by means of which we can discriminate cases in which the goal is achieved from others in which it is not? It seems that answers one can find by looking at primary sources

[1] On a similar basis, one can argue that to set up a term definition has different philosophical implications from carrying out a proof: the latter for Aristotle also comprises a declaration of existence of the "object" of it (see [7, p. 141]).

[2] For a rather exhaustive picture of the quite complicated issue of definitions in classical Greek philosophy, the reader is referred to the relevant chapters of [9] that, as it might be clear already, I have used myself to obtain the crucial information I gathered for the sake of this introductory section.

from the same period of time in which this view was proposed are inconclusive, and not because these issues were uninsteresting. On the contrary, scholars observe quite an activity in the attempt of sistematizing the search for good definitions[3]. Yet, there seems to be lack of a clear conclusion in the end of it. Therefore, it seems legitimate to look at Aristotle's attempt to achieve some sort of characterization of good definitions as the latter was carried out in his *Topoi* (see [8, p. 226]), as paradigmatic of a widely shared difficulty in this respect. According to him, as a matter of fact, a good definition should: (1) be a universal and true predication of its definiendum; (2) should put the latter in its *genus* and *differentia*; (3) should be the proper account for its definiendum; (4) should state the essence of it; (5) should define it well. If the problem was to find a norm, that is to determine where lies the difference between good and bad definitions, then Aristotle's characterization is clearly disappointing. As I said, it is the sign of a widely shared difficulty that seems to be tied up with the strict connection between definitions and essence that this tradition aims to devise. This difficulty is precisely the point of my departure. Not because I aim willing to pursue this path any further. Rather, because I will proceed in a contrary direction. As the title of this paper suggests, I will focus on a special kind of definitions, namely circular definitions. A definition is circular in the sense of this paper in case the term to be defined literally occurs in the property defining it. As far as I understand the view I have spoken of so far, there is no way to reconcile it with admitting that definitions might be of this latter type: for, if one wills to genuinly capture the essence of something, then how can this essence be presupposed already without trivializing the whole attempt? This means, as I said, that in order to pursue my goal I have to depart from the classical view of definitions in the first place, and find some alternative framing perspective that is coherent with the point I want to make.

2 The form of definitions

In his his 1810 work [2, ch. 1, §8], Bernard Bolzano defines mathematics as a "*science which deals with the general laws* (*forms*) *to which things must conform in their existence*", and he adds that "[b]y the word "*things*' I understand here not merely those which possess an *objective existence* independent of our awareness, but also those which simply exist among our *ideas*, either as *individuals* (i.e. *intuitions*), or simply as *general concepts*, in other words, *everything at all which can be an object of our perception*". On the one hand, Bolzano assigns in

[3]Plato's "method of division", based upon going through sequences of opposite features, for instance, is presented as marking the difference between the early dialogues of his, where Socrates attempt in solving quests end in *aporia*, and late dialogues where the solution of the problem originating the discussion is more frequent (see, for instance, [3] and [11]). Aristotle's dialecthic is similarly seen as an attempt to examine the problem of predication (which is one of the critical issues of definitions, philosophically speaking) via questions about opposites (see [8]). In addition, it is suggested that the relationship between definitions and proofs in Aristotle might be a source useful to clarify in which cases one can claim that the essence of things has been isolated, i.e. good definitions have been set up (see [6, §3, in particular]).

this way to mathematics the role of a universal science, a *mathesis universalis*, deputed to isolate the laws of existence of things. On the other hand, he acknowledges that the domain of application of this science is basically unlimited, and comprises both elements belonging to the outer, objective reality, as well as items belonging to the intellectual domain. Bolzano's phrasing seems to be also suggesting that the basic difference between the two sorts of things to which the laws of mathematics apply, is that while those belonging to the former domain are independent of our intellectual activity and therefore come along with characters of their own, those belonging to the latter domain instead may be influenced by one's mental disposition instead and be subject to one's personal inclinations.

Bolzano's view can be regarded as the sign of change of paradigm that has affected also the way philosophical research is also pursued, and which, beside him, has been exploited by many others in the subsequent decades, and up to the present-day. This paradigm ties the philosophical investigation to the mathematical and to the formal logical one. What if this paradigm is applied to the case of definitions? The aim of this section is to restart the analysis of the topic on this basis, and draw some of the consequences that follow from this change of attitude.

The first thing I will do is to sharpen the object of investigation. I will follow a concrete, formal approach. In particular, by "definition" I will refer to a linguistic expression that is made out of three parts:

1. a left–hand component, commonly known as *definiendum*, which contains the term to be defined;
2. a right–hand component, the *definiens*, which is the condition defining it;
3. a middle component, which I will refer to as the *definitional formula*, that connects the other two parts.

I will also assume that definitions which are of interest here come in the following "standard form" accordingly:

$$x \text{ is } P \equiv \Phi$$

where the left–hand side of it, "x is P", is the definiendum (P being the term to be defined), the right–hand side Φ is an expression of the language that plays the role of definiens, and $\equiv$ is the definitional formula (which, before I say something more precise on that, can be read as "is defined as", or "is defined by"). To assume that definitions always occur in one form is of course a simplification of the actual case in which they come in many forms instead, like:

- "Man is rational animal";
- "Courageous means brave";

- "Any given $x \in \mathbb{N}$ is prime if and only if x is greater than 1 and has no positive divisors other than 1 and x itself".

To justify the reduction of definitions to standard form is easy for examples such as the first and the last in this list (at least, as long one accepts a set-theoretic view of properties which has become common nowadays). It may require some more labour for cases like the second, which is where one is confronted with a definition of a special kind, i.e. a "dictionary" definition by means of which one fixes the meaning of a term. Since it lies outside the goal of the present paper to classify definitions according to the role they play, I am just skipping any further comment on that and stick to the assumption that definitions I speak of are in the above standard form as far as the left–hand side of them is concerned[4].

To consider definitions playing different roles like those I have displayed, is useful to justify the use of the symbol $\equiv$ for the definitional formula instead. To avoid specifying the definitional formula is my attempt to capture the variety of formulations it may actually take. About this aspect, however, I am not going to say anything more than that, not at this stage at least. There is some specific issue regarding the definitional formula in connection with the transition to the formal level I will be dealing with later on in the paper. Therefore, I will engage in a more detailed discussion of this feature at that stage.

Something preliminary should be said instead concerning the relation between the left–hand side and the right–hand side of definitions. In particular, my main concern here is whether they belong to the same portion of the language or not. The problem is notorious, but I will briefly try to justify the issue here for the sake of self–containedness. Roughly speaking, the whole matter stems from the clash of two views. In the one, definitions are tools for assigning new terms their meaning, therefore for making them "usable" to users of the language they will belong to henceforth. According to this idea, if someone is making use of a certain language $\mathcal{L}$ of which Φ is a legitimate expression, the above definition makes available the new term P, therefore upgrades language $\mathcal{L}$ to language $\mathcal{L}^+ := \mathcal{L} \cup \{P\}$.

Albeit natural, this view of languages seems quite artificial if confronted to the way things proceed in the actual situation. As far as the "growth" of languages is concerned, things seem to be going less smoothly and appear to be much more complicated in real life. Although new terms defined by means of pre-existing linguistic resources of course occur, it may also happen that old terms get new meanings in addition to those they have already. Influences between different languages are also possible, and it is not rare to see foreign words being used by other mother-language speakers. A dictionary is a good example of this more complicated way in which languages gain new expressive power. It is certainly true that every word has its own definition of meaning. However, the analysis of the relationship between each term and its own defining condition is a matter which is influenced by many different factors. This makes it

[4]The reader interested to some more elaborate view in this respect, can be referred to the quite comprehensive taxonomy provided by Anil Gupta in [14].

possible, if not likely, that the relationship between the defining condition Φ and the term P it defines be more complex than someone is willing to consent. Just to make a concrete example in this respect[5], the current (January 2018) online version of the Merriam-Webster dictionary defines a *hill* "a usually rounded natural elevation of land lower than a mountain", and a *mountain* "a landmass that projects conspicuously above its surroundings and is higher than a hill". It should be clear that the two definitions create a circle that makes difficult to assign a definite meaning to the terms involved therein (for, one is required to know the meaning of "mountain" to understand the definition of "hill", but the meaning of the latter is required to get the meaning of the former also).

The extremest case of this more complicated relation between a definiendum and its own definiens, is represented by a term P that occurs in its defining condition Φ. Is this blatantly circular case also supported by evidences? In what follows I take up the issue, arguing that a positive answer to this question has a certain degree of plausibility, and leads to some interesting consequences if addressed at the formal level.

3 Why bother with circular concepts

Let

$$x \text{ is } Q \equiv \Psi(x)$$

be a standard definition of Q in a language $\mathcal{L}$, where the definiens Ψ belongs to $\mathcal{L}^{-Q} := \mathcal{L} \setminus \{Q\}$, therefore is free of occurrences of the definiendum. If one assumes that the semantics for the language $\mathcal{L}^{-Q}$ is fully developed in the form of a model $\mathbf{M}$, there are clear reasons for arguing that this definition is indeed a good one. On the one hand, it allows to provide Q with a meaning, the extension of its application in $\mathbf{M}$, in the form of the collection:

$$\{a \in |\mathbf{M}| : \mathbf{M} \models \Psi(\overline{a})\}$$

(where $|\mathbf{M}|$ is the domain of $\mathbf{M}$ and, for every element a of it, $\overline{a}$ is the "name" for it in $\mathcal{L}^{-Q}$)[6].

On the other hand, it provides Q with some clear rules of use, a natural logic, in the form of rules for introduction and elimination of the definiendum:

$$\frac{\Psi(\overline{a}) \text{ holds in } \mathbf{M}}{Q(\overline{a}) \text{ holds in } \mathbf{M}} \ (Q \text{ in}) \qquad \frac{Q(\overline{a}) \text{ holds in } \mathbf{M}}{\Psi(\overline{a}) \text{ holds in } \mathbf{M}} \ (Q \text{ out})$$

With respect to the traditional approach, this marks a first advantage of assuming the mathematical approach I took, following Bolzano's quote: if one leaves aside considerations about definitions capturing the essence of things, and if one is willing to assume a more concrete attitude toward the topic, it is possible

[5] Timo Beringer is responsible for having spotted and make this example of actual circularity available to me.

[6] As is well known, we can assume that $\mathcal{L}^{-Q}$ has names for every element of $|\mathbf{M}|$ in all cases that are relevant to the present discussion without loss of generality.

indeed to set up a credible paradigm of how definitions should look like. Now, granted that this is the paradigm to refer to, then there seems to be no point in considering the circular cases mentioned above. For, circular definitions make it hopeless to try obtaining either of its two characteristic features, as neither an extension, nor a usable logic can be achieved. This seems to be the quite obvious and unavoidable conclusion. But, is it really so?

3.1 Practicing circularity

The example extracted from the Merriam-Webster dictionary at the end of §2 is a real-life counterpart of the formal case in which the definienses of two terms, say P and P', refer one to other in the sense that the definiens Φ of P contains occurrences of the term P', and the definiens Φ' of P' contains occurrences of P. Despite the actuality of it, this case might not be enough to convince that also the case of a term P occurring in its own definiens Φ, which is more radical as I suggested, should be regarded as legitimate. However, closer inspection at how some common linguistic expressions are used, may lead to that conclusion.

Suppose that someone has to verify that the length of a certain object corresponds to a given, fixed amount. Let us say, to make the example more precise, that one needs to check that the edge of a table has length one meter. Then, he, or she, would follow the well–known procedure: take a ruler (which we assume has exactly the length to check – one meter in this case), let one edge of the ruler correspond to one edge of the table, and verify that the opposite edge of the ruler precisely correspond to the opposite end–point of the side of the table under scrutiny. If this happens, then the conclusion would be that the edge of the table and the ruler "have the same measure", therefore that the former measures one meter like the latter. The story suggests that one is implicitly referring to a general definition which reads:

$$x \text{ measure one meter} \quad \equiv \quad \begin{array}{l}\text{there exists } y \text{ which measures one meter,} \\ \text{and } x \text{ has the same measure as } y\end{array}$$

Now, the definition is clearly circular, and its definiens "captures" the very process used to verify whether the specified property applies or not (that is, to say that it holds true of a given object x corresponds to successfully applying the previous verification procedure). This tells us something more about the situation we are considering, as not only it seems we have found a real–life example of circular definition in which the definiens refers to the very same definiendum it defines, but the circularity of it is not even vicious: rather, it virtuosly contributes to practical decisions concerning the fact that the attribute it involves legitimately belongs, or belongs not to the object considered. Notice that I am not claiming that the said attribute *is* circular. Having decided to take distance from considerations about definitions being somehow involved into disclosing the essence of things, this is not the goal of the observation. What I am emphasizing here is that, under very common circumstances, we act *as if* this attribute was circular, granted that the verification process we make use of reflects this feature of the definition, and legitimates it. This seems to

me to be enough to conclude that circularity, even in its most extreme form, is part of real life. It also proves that it behaves well, i.e. it does not lead to inconsistencies as one might be brought to think. As a matter of fact, even more is true. It seems to me that the circular feature affecting properties referring to exact measures is inherited by properties which refer to measures indirectly. Think of the well–known controversial concept of "heap". Due to the paradox it is notoriously prone to, it turns out that it is impossible to reach a categorical definition of it by identifying the corresponding property with an exact measure of "unities", like, for instance, grains of sand: for, any of such number being fixed, our intuition suggests that a collection of grains where only one of them has been taken out should be eligible to be called "heap" anyway, then causing the property to disappear, grain after grain. Now, why not looking at this matter in the light of what we have just concluded, and assume that when we call something "a heap" we are not assigning this property by categorical judgments, but via circular reasoning instead? The above difficulty about reaching a final decision on what should be called "heap" and what should not, may suggest that something like the following definition is closer to what leads us to draw conclusions in this respect:

$$x \text{ is a heap} \quad \equiv \quad \begin{array}{l}\text{there exists } y \text{ which is a heap, and } x \text{ has} \\ \text{the same measure as } y\end{array}$$

A similar approach would work in case other vague predicates are considered, like "tall", "small", "big", etc. Notice that these also contain an indirect reference to measuring.

It is not my objective, however, to make a case here about this matter[7]. The above observations seem to me to justify a more modest conclusion about circularity being, in some form or another, an actually existent feature of our mode of thinking in the everyday practice. In turn, these real-life examples suggest that despite differences which are critical to both the issue of providing the defined concept with an extension, and of describing a natural logic of it, circular definitions are "handled" somehow, or, at least, that they are not stumbling blocks to the practical purposes the said concepts are used for. This conclusion is strengthened by the consideration of some more examples, which appear to be more "sophisticated" than those I have been considering so far (in a sense that I will try to argue for below), but which help strengthening this conclusion.

3.2 On truth, and other circular concepts

In his seminal paper on the concept of truth in formalized languages [19], Alfred Tarski pursued a systematic study of this notion starting from the intuition of it that is most naturally related to the view according to which: (i) languages are tools for speaking of reality, and (ii) truth is the property of statements

[7] See [1] for a full proposal about how to deal with vagueness by means related to those I refer to in §4 to cope with circular definitions.

that follows from the agreement between what is said of reality and how reality is. As a result of his analysis, Tarski was lead to isolate the crucial principle comprising this intuition, which takes the form, for every statement φ of the chosen object language:

$$\text{“}\varphi\text{” is true if and only if } \varphi$$

The principle is supposed to express the idea that truth stems from the said relationship between what one asserts and how “things” are. This is done by comparing linguistic objects (on the left) with their “propositional content” (on the right), that is, a statement φ with what it “says” (independently of *how* φ says it – for instance, independently of how φ is formulated according to the language it belongs to). This latter content also corresponds to a state of affairs, a condition that takes place in the “world” on which truth of φ ends up depending upon. Now, it is quite clear that the principle above is *not* a definition of truth itself in Tarski's view. Rather, the derivability of all instances of it for a chosen language is part of what Tarski calls the *convention T*, the “test” that certifies that a certain definition of truth for that language is adequate.

Tarski's idea has inspired some more investigations related to a deflationary stance on truth, which goes back to the work of many (like Frege, Ramsey, Quine, to mention a few) who favoured an anti–metaphisical approach to this notion. With new champions of such view recently rearing up their heads, the above Tarskian principle was brought back in the spotlight (assuming it had ever been left behind), and the collection of all biconditionals corresponding to it is now usually taken to express everything that can ever be said on truth according to deflationists. In absence of an explicit definition of truth, this has lead someone to stress the fact that, at some fairly natural conditions, the collection of these biconditionals itself yields a (circular) definition of it[8]. Let $\mathcal{L}$ be a formal language that contains a dedicated predicate $T(x)$ to express that “x is a true sentence of $\mathcal{L}$”. Suppose that $\mathcal{L}$ has a citation device “$\cdot$” for calling its own statements “by name”. For instance, this holds if $\mathcal{L}$ extends the formal language of arithmetic $\mathcal{L}_{PA}$, as the latter guarantees the existence of a name $\#\varphi$ for every statement φ through application of the standard gödelization technique of assigning to any statement φ of $\mathcal{L}$ a unique natural number n_φ as its numerical code, and then to set $\#\varphi$ as the *numeral* corresponding to it (i.e., $\#\varphi = succ^{n_\varphi}(\overline{0})$, where the latter indicates the n_φ–ary application of the successor operator of $\mathcal{L}_{PA}$ to the term $\overline{0}$, which is the term for the number zero). The above principle for such an $\mathcal{L}$ would then take the form:

$$T(\#\varphi) \equiv \varphi$$

If read as a definition, then $T(\#\varphi)$ would act as a definiendum, φ would be its definiens, and $\equiv$ the definitional formula. The latter would be open to a two-fold interpretation: a semantical interpretation according to which, for every model $\mathbf{M}$ of $\mathcal{L}$, the left–hand side of the definition holds in $\mathbf{M}$ just in case the right–hand side does, and a syntactical interpretation, which would

[8]This view has been supported, for instance, by Gupta and Belnap [15].

be made possible by the presence in the language $\mathcal{L}$ itself of a biconditional connective to represent the definitional formula. If this biconditional is the one that stems from the usual material implication, this interpretation in an appropriate deductive setting would be the same as assuming that the left–hand side of the definition is provable just in case the right–hand side is.

Since I will come back to the relationship between these two interpretations in more general terms later, there is no point to discuss it with respect to this special case here. Some more delucidation is required instead, as to how this connects with circularity. For, the peculiarity of the case under scrutiny is that the above principle is schematic, i.e. indicates a multiplicity of instances which, taken as a whole, would provide us with a definition of the concept of truth for $\mathcal{L}$. So, the feature I have taken so far to be the mark of circularity, namely the presence of the definiendum in the definiens, would be true actually only of those statements φ which feature occurrences of T (i.e., do not belong to the sublanguage $\mathcal{L} \setminus \{T\}$ of $\mathcal{L}$). However, the mere fact that a sentence φ is of this latter sort, would be not enough to conclude that the corresponding instance of the above schema is circular in the said sense: assume that the occurrences of the definiendum T might be "dispensed with" indeed, for instance because of the existence of a statement ψ of $\mathcal{L} \setminus \{T\}$ which is logically equivalent to φ; this would legitimate the substitution of ψ for φ in the relevant instance of the above scheme, and the consequent elimination of circularity from it. The point is precisely that not all occurrences of T in sentences of $\mathcal{L}$ are eliminable in this sense, as there are cases in which the presence of the definiendum in the sentence providing the definiens is unavoidable, as no logical equivalent of it can ever be found among formulas of $\mathcal{L} \setminus \{T\}$. This would be the case, for instance, of the sentence $\forall x(T(x) \vee \neg T(x))$. The previous schema would yield the following instance in this case:

$$T(\#\forall x(T(x) \vee \neg T(x))) \equiv \forall x(T(x) \vee \neg T(x))$$

Since the sentence on the right is a universal, the claim that truth is eliminable from the definiens in this case amounts to say that is eliminable from every instance of it. However, it should be clear that this is not possible for at least the instance of the definiens which is obtained by substituting x with the definiens itself, yielding

$$T(\#\forall x(T(x) \vee \neg T(x))) \vee \neg T(\#\forall x(T(x) \vee \neg T(x)))$$

(for, no further attempt to make all of the occurrences of T disappear would work). Hence, if the collection of all instances of the biconditional under scrutiny is taken as a definition of truth, then this leads to a circular definition in the sense of this paper. With respect to the cases of circularity I have discussed earlier in §3.1, however, this case is of a different nature: in order to ascertain that truth is circularly defined by the above principle, it is required to pass through the consideration of a formal language, or to assume that the language one is considering features a citation device for sentences with the required properties, let alone the fact that one has to accept the idea that the collection of formulas

I have been speaking of so far provides us with a definition of truth for the chosen language. Therefore, this example comes with a degree of sophistication which is higher than the examples I spoke of previously. This is not to say that is useless for the goal I pursue. On the contrary, it similarly reveals how some apparently very natural intuitions of ours (those on which the idea of truth as correspondence of language with "facts" is based), may end up in a circular definition once they are made precise. Also, albeit being the most known and the most cited case when circularity comes into play, this is not even the only example one can think of in this respect. Another one comes from a totally different kind of situations.

Suppose we are looking at a group of agents playing and, given a certain state of the play, we would like to determine what the players should do next, or which one is the rational action for them to play among those they have available. The problem is less specific than it may seem: it is well-known that a variety of real-life situations can be described in a game-like fashion. To make things easier, let us first consider a very simple case: Alice is looking for a part-time job that could ensure her to earn some money while she is finishing her studies at university; she then finds this offer from a company which is seeking for employees who may increase the number of "likes" on websites of clients over their competitors; Alice decides to accept the offer, and, according to the contract she signs, she will be paid more, the more "likes" she places. If we assume that no legal, moral, or other forms of infringements are at risk of limiting Alice's activity, then the situation is so simple that there seems to be no possible doubt about what Alice should do, or, to reconcile the example with the problem we started from, about what is the rational action for Alice: she is expected to place as many "likes" as she can during worktime, and therefore get the highest pay she can get (thereby, as game–theorists would say, maximizing her utility). The problem now is to clarify whether the intuition which works well in this case is strong enough to survive to more complicated situations and be used to set up a definition of the concept of "rational choice". So, assume that Alice's company is expanding and provides her with a team mate, a new employee named Bob, but at the same time changes the clauses of Alice's contract to the effect that now she (and Bob) will be paid more in all situations in which their actions agree, that is when they both like or dislike a website, while they will be paid less whenever they disagree. The strategy Alice should apply in the new scenario is clear as well: she is expected to "play" *like* whenever Bob does that, and similarly play *dislike* if Bob is doing the same. The same is for Bob of course, which brings us to the following moral: Alice's rational action is the reply to Bob's rational action that entitles her to get the highest pay. If this moral was tentatively extracted from the situation and given the form of a general principle, of a definition, then it would read more or less as follows:

$$\begin{array}{rcl} \text{An action } x \text{ is rational} & \equiv & \text{there exists a rational action } y, \text{ and performing} \\ & & x \text{ against } y \text{ maximizes the performer's utility} \end{array}$$

But, if this were really taken as a definition of the concept of "rational action", then we would have to admit that this case is not different from the

case of "true" for a proposition, and does not even seem to be far from the concept of "heap" or the other concepts from the same family I mentioned in the previous part of the section. Of course, there is a fundamental difference between the analysis of the concepts of "truth" and "rationality" on the one hand, and the other examples I analyzed earlier. This difference is what I meant to emphasize by referring to the latter cases as "real-life" concepts, and call the others "sophisticated" instead. While in the cases I analyzed first circularity is assumed to be part of our daily practice with the related concepts, both in the case of truth and rationality there is an undeniable level of abstraction from reality to be accepted, that leaves open the door to criticisms about faithfulness to our actual use of these notions. "Truth" and "rational action" could be circular at this higher level, without being "really" as such. In other words, circularity here could be considered as a side-effect of the *in vitro* approach we are pursuing to make sense of the *in vivo* situation. I avoid taking a stance on this, because it seems to me that the two parts of this section taken together provide enough support to the view I intended to convey: circular definitions affect our life, both the "real" and the "sophisticated" life of ours, and therefore they require a specific justification that could let us better understand why they are sometimes regarded as legitimate as ordinary definitions are.

4 Hypotheses and their revision

Let us go back to the measuring example of §3.1. That was, I claimed, a case where a circular definition is unproblematic since it reflects the very process by means of which the defined property is ascribed or not in actual cases. If this is so, then a deeper analysis of the situation may give important indications for solving the problems with circularly defined concepts in general. Now, the very process the definition in question refers to, can be briefly described as follows: in the case of two objects a and b, having determined that a measures the same as b allows me to conclude that a measures one meter (in short: $1M(a)$), since I know, or, better, I have assumed that $1M(b)$ holds first. The said procedure is three-fold: (i) starts with the hypothesis that the property under scrutiny holds of b (that is, that $1M(b)$ is the case), (ii) proceeds by the verification that a satisfies the condition defining it, for which it is enough to verify that a has the same measure as b granted that $1M(b)$ holds by hypothesis, and (iii) ends with the conclusion that $1M(a)$ is the case as a consequence. This seems to be enough to flag a first general consideration: circular definitions do not allow to assign to the definiendum an extension categorically, but they do it hypothetically (see also [12] on this). To make the observation more precise, let us assume to be working with a first-order formal language $\mathcal{L}$, as before, with model $\mathbf{M}$ and domain of individuals $|\mathbf{M}|$, and let us also assume to have a unary predicate P of $\mathcal{L}$ circularly defined by:

$$P(x) \equiv \varphi(x, P)$$

(where $\varphi(x, P)$ is a formula of $\mathcal{L}$ that features occurrences of both x and

P as explicitly indicated). Then, the hypothetical character of the previous reasoning with circular concepts amounts to the passage from collection H to collection H' below:

$$H = \{x \in |\mathbf{M}| : P(\overline{x})\} \mapsto H' = \{x \in |\mathbf{M}| : (\mathbf{M}, H) \models \varphi(\overline{x}, P)\}$$

(where $(\mathbf{M}, H) \models \varphi(x, P)$ indicates the obvious modification of the usual validity relation in which H is used to evaluate occurrences of formulas of the form $P(s)/\neg P(s)$ in φ - see footnote 11 below for details -, and, for every $x \in |\mathbf{M}|$, $\overline{x}$ is the term of $\mathcal{L}$ that refers to it). More precisely, it corresponds to the passage from the hypothesis that b belongs to H to the conclusion that a belongs to H' then. The latter, which provides P with a (possibly) new and refined extension, acts as the *revision* of the starting hypothesis.

Granted that, the full import of the difference between non-circular definitions and circular ones amounts of course to the fact that, if the conclusion about an object possessing a certain property is not categorical but hypothetical instead, then it is not stable in time and is subject to change. So, in our running example it cannot be excluded that there is another object c such that, if $1M(c)$ is supposed to hold, then $\neg 1M(b)$, as well as $\neg 1M(a)$, turn out to be the case. Therefore, if one puts the starting hypothesis, its revision, the revision of this revision, and so on in a "stream", so to form a *sequence* $(H_i)_{i \in I}$, where $(I, \prec_I)$ is an index set totally ordered by $\prec_I$, then it may happen that a given object d which does belong to a set H_n in the sequence, does not belong to a subsequent set H_m with $n \prec_I m$, and that it belongs again to a set H_p coming later on (that is, such that $m \prec_I p$). As one should say in mathematical terms, the revision sequence $(H_i)_{i \in I}$ is *non–monotonic* with respect to the subset relation as $H_i \subseteq H_j$ might not be always the case for $i \prec_I j$.

Indexed sequences of sets have been studied at length mathematically speaking, for instance in the form generalized inductive definitions of both monotonic and nonmonotonic character (see, for example, the overview [10] and the literature cited there). One of the motivating feature for analyzing sequences is the existence of a *closure set*, a point in the sequence which can be regarded as the natural "end" of it. In the most fortunate case, closure sets may take the form of *fixpoints*, i.e., for a given sequence of sets $(X_i)_{i \in I}$, the form of a set X_j such that $X_j = X_h$ holds for every $h \in I$ such that $j \prec_I h$. As is well known, the existence of fixpoints is related to features of the sequence which, in cases like the one we are looking at, also correspond to logical properties of the definiens which is responsible for the revision step[9]. Due to the general approach to circularity we are pursuing here, the existence of fixpoints for all revision sequences is not granted. Therefore, it becomes natural to ask: what

[9]The key result related to the existence of fixpoints for an infinite sequence of sets is the theorem by Tarski and Knaster according to which every monotone sequence admits fixpoints. The result extends to sequences of hypotheses determined by circular definitions in the sense of this paper, whenever the definiendum occurs only positively in the definiens (an occurrence of a predicate P in the logical formula $\varphi(x, P)$ being positive if P occurs in the scope of an even number of negations).

counts as an "end" of a revision sequence, and how long can we expect to keep revising hypotheses before coming to it?

Despite the overall nonmonotonic character of the sequence they belong to, not all hypotheses reflect this feature of it as one can think of a number of situations in which a set of the sequence might be regarded as possessing some form or another of "local stability", even without being fixpoint. For instance, to mention just a few cases that may help the reader get the idea, this might be said of a set H_i for which there exists $j \in I$ with $i \prec_I j$ such that $H_i = H_j$; or, of a set H_h such that, for every $i \in I$ with $h \prec_I i$ there exists $j \in I$ with $i \prec_I j$ for which $H_h = H_j$ is the case; or also of a set H_h such that there exists $i, j \in I$ with $h \prec_I i \prec_I j$ and, for every $k \in I$ with $i \prec_I k \prec_I j$, $H_h = H_k$ is the case. Being all these examples of situations in which the global instability related to the nonmonotonic character of the sequence somehow "breaks down" at the local level, and being it possibile to present them as natural modifications of the ordinary fixpoint case, they can all be appealed to in order to single out "solutions" of sequences of sets produced by the revision machinery described above.

The point is whether they are feasible solutions, that is how likely one can expect to have indices like i, j and k above in (possibly all) actual cases, and how long must the revision process go along before getting them. The answer to the first issue is positive (see [15]), but it turns out that the answer to the second one is much more interesting. As a matter of fact, this is the point where revision sequences split into *finite*, and *infinite* ones.

Finite revision sequences are cases in which you do not have to wait for long, so to say, to reach a solution you would call "stable". As a matter of fact, it turns out that finite revision sequences are characterized by a very peculiar structure: they consist of an initial segment which contains different hypotheses, followed by blocks of other hypotheses which keep repeting themselves. It is easy to explain this structure in more precise terms by taking $(\mathbb{N}, <)$, where $<$ is the ordinary "less then" relation over $\mathbb{N}$, for $(I, \prec_I)$. Then, a finite revision sequence is a sequence $(H_n)_{n\in\mathbb{N}}$ for which there exists $k \in \mathbb{N}$ and $p_k \in \mathbb{N}$, the *period* of the sequence, such that $H_{m+p_k} = H_m$ for every $m \in \mathbb{N}$ with $k \leq m$.

To reach a similar situation in the case of infinite sequences is a more complicated matter. In particular, an infinite sequence is such that ω-long subsequences of it do not feature periodic repetitions of blocks of hypotheses, and it is therefore required to extend the sequence transfinitely. So, the index set $(I, \prec_i)$ must be order-isomorphic to an appropriate initial segment of the set of ordinal numbers On containing ω, and a rule must be set to state how elements H_l's of the sequence are formed for indices l corresponding to limit ordinals. Several proposals have been advanced in this respect (see [15], again). One that is particularly easy to understand, and which can serve the purpose of providing the reader with a working example, is based upon making use of the liminf intuition according to which any such set H_l contains the elements of the previous sequence of hypotheses which are candidate to be stable, in the sense that they belong to every previous sets from a certain point onwards (formally: $x \in H_l$ if and only if there exists $i \prec_I l$ such that $x \in H_j$ for every $i \prec_I j \prec_I l$). By

cofinality arguments, then one proves that for revision sequences in this transfinite form there are (denumerable) closure ordinals, i.e. hypotheses H_s which contain the stable elements of the sequence as a whole (that is, $x \in H_s$ if and only if there exists $i \in I$ such that $x \in H_j$ for every $i \prec_I j$ – see, e.g., [15] for details.).

If analyzed according to the revision-of-hypothesis intuition, circular definitions split into two classes accordingly: *finite circular definitions*, which are those cases whose definiens will give raise to revision sequences which become "regular" in finitely many steps, and *infinite circular definitions* which comprise all cases of definienses that require the whole revision process to be iterated transfinitely to find regular hypotheses instead. So, for instance, the concept of rational choice mentioned in §3.2 is an example of the former definitions (see [13]), while the concept of truth for formalized languages is an example of the latter ones. What interests us here, however, is to establish whether the machinery sketchly introduced here can be used to try solving the issues that affect circular definitions in general, and therefore used to fill the gap with the non-circular cases.

5 The extension and the logic of circular concepts

The process of assigning a hypothetical extension to a circular predicate and then revise it, can be presented (as I have tried to do above) as the most natural modification of the usual approach to definienda in non-circular definitions. The existence of regularities and "closure stages" in revision sequences, is the mark that constructions of this sort are a promising approach to use as far as the goal of making sense of circular definitions is concerned, despite their nonmonotonic character. The question is whether they also provide solution to the issues that make the circular case an exception to what can be done in the non-circular one.

The first of those issues concerns the missing extension for circular predicates. One important feature of revision sequences the reader might have noticed already, is that any revision sequence produced out of a given circular definition is relative to the hypothesis one chooses to start from. Different initial hypotheses will induce different sequences, which will feature different periods and different repeting hypotheses. To make up for that, there are two major solutions. The first one consists in letting all revision sequences start from one and the same initial hypothesis. A choice in this respect can be done coherently with the spirit of the revision–theoretic approach as a whole. As a matter of fact, the latter is based upon the idea that circular definitions are dealt with by means of a trial-and-error procedure, starting from an inital guess that gets revised and refined as the process goes along, eventually reaching some stable conclusions. This view may motivate the idea that at the beginning of a revision sequence should stay the null hypothesis, the empty set, that would correspond

to the very initial level where no guesses at all are made about what the circular predicate holds of, and which represents the original state of ignorance one is gradually lifted out of until some more stable hint is achieved eventually.

A different solution can be justified by assuming the quite opposite viewpoint according to which circular definitions give no indication that a starting hypothesis should be preferred to another. The bootstrapping procedure of hypotheses and revisions is justified by the fundamental character of circular predicates, but it provides us with no more hints than that. Therefore, any guess is admissible, and none of them should be regarded as "canonical". If all revision sequences originated by one and the same circular definition are equally legitimate as a consequence, then one should rather look at what happens in all of them and set the extension of the circular predicate to contain just the elements behaving "stably" everywhere in this larger sense (i.e., set it to coincide with the intersection of repeting hypotheses from all revision sequences one can think of in the finite case, and with the interesction of all closure hypotheses in the infinite one).

It should be clear that, these two proposals notwithstanding, the prospect of extracting some categorical information about circular predicates from the revision–theoretic construction remains controversial[10], and, one may argue by the way, could it be differently? For, one could think that the construction itself remains the only moral of the whole story: the distinctive feature of circular predicates is the bootstrapping procedure itself as I said already, and the main character it expresses is the effect of the hypothetical reasoning it is based upon; therefore, any attempt to draw a categorical conclusion is bound to contradicting its own nature.

What about the other feature we were wondering about, the logic of circular definition? If the construction based upon revising hypotheses is what can be justified best by the living examples of circularly defined concepts, then to understand the logic behind them is to disclose what logical properties are entailed by the construction itself. It turns out that this can be done again by suitably modifying what occurs in the non-circular case in a natural way.

At the beginning of §3, I suggested that the existence of a semantics, which provides us with a sense of what "holds" and what "holds not", gives the ground for the basic rules (Q in) and (Q out) if Q is a non-circular predicate. Let us assume again, that P is circularly defined within a formal language $\mathcal{L}$ as before. That is:

$$x \text{ is } P \equiv \varphi(x, P)$$

The approach via hypotheses and revisions can be presented as filling the missing details about the semantics of circular definitions, since a revision sequence is built up by extending the usual validity relation $\models$ for formulas of a given language[11]. Due to the peculiar structure it features, a revision sequence

[10] See, for instance, [18], [17] and [12] for a lengthy discussion on this topic.

[11] To be more precise, provided that a model $\mathbf{M}$ for $\mathcal{L}^{-P}$ is given, then one inductively defines $(\mathbf{M}, H_i) \models \theta$ for every hypothesis H, $i \in I$ index set, and θ from $\mathcal{L}$, as the smallest

$(H_i)_{i\in I}$ allows us to formulate judgements about formulas of $\mathcal{L}$ as holding or not holding at a certain stage in the sequence. In particular, the way the revision step is conceived legitimates the following modification of the definiendum introduction, and elimination in the case of a circular predicate:

$$\frac{\varphi(\overline{a}, P) \text{ holds at } i}{P(\overline{a}) \text{ holds at } i+1}\ (P \text{ in})_i \qquad \frac{P(\overline{a}) \text{ holds at } i+1}{\varphi(\overline{a}, P) \text{ holds at } i}\ (P \text{ out})_i$$

(where $i+1$ indicates the immediate successor of i in I[12]).

Now, rules for definiendum introduction and elimination (Q in), (Q out) in the case of non–circular concepts justify the introduction of axioms of the form:

$$Q(x) \equiv \Psi(x)$$

where here $\equiv$ represents the usual biconditional connective. The transition is made legitimate by the properties of material implication that allow to internalize deductive steps, and to pass, for instance, from a derivation of a formula θ from ϕ in a system of axiom S that is adequate to the given semantics, to the derivation of the formula $\phi \supset \theta$ in that system[13]. As the backward semantics changes to make sense of circular definitions, the corresponding transition in the case of P is no more justified. As a matter of fact, the rules $(P \text{ in})_i$, $(P \text{ out})_i$ above hide the conditional passages summarized by: "If $\varphi(\overline{a}, P)$ holds at stage i (of the revision sequence $(H_i)_{i\in I}$), then $P(\overline{a})$ holds at stage $i+1$", and "If $P(\overline{a})$ holds at stage $i+1$, then $\varphi(\overline{a}, P)$ holds at stage i" respectively. Actually, as far as these rules are concerned, the conditional reasoning they encompass can be simplified by avoiding any reference to stage labels, as it amounts to the following statements:

"If $\varphi(\overline{a}, P)$ holds at the current stage, then $P(\overline{a})$ holds at the next one"
"If $P(\overline{a})$ holds at the next stage, then $\varphi(\overline{a}, P)$ holds at the current one"

The simplification makes clear that we are seeking for a conditional connective that may convey this sense of passing from one stage in a revision sequence to the next one, and back. Clearly, this cannot be achieved by means of material implication, which is justified differently. Gupta and Standefer [16] proposes new conditionals that are specifically designed to encapsulate the required reasoning. These conditionals, which they refer to as the "step conditionals", are therefore candidates to provide us with the logic which, in the spirit of Bolzano's quote from §2, may let us conclude that we do have a proper understanding of circular definitions. This is the goal I take up in the next section.

---

relation that extends validity in $\mathbf{M}$ for sentences of $\mathcal{L}^{-P}$, by interpreting occurrences of the form $P(\overline{a})$, $\neg P(\overline{a})$ according to H_i (in particular, by putting: $(\mathbf{M}, H_i) \models P(\overline{a})$ if and only if $a \in H_i$, and $(\mathbf{M}, H_i) \models \neg P(\overline{a})$ if and only if $a \notin H_i$).

[12] That is, $i \prec_I i+1$, and for every $j \in I$ such that $i \prec_I j$, then $i+1 \preceq_I j$ holds.

[13] I am using here the symbol $\supset$ for material implication, saving the usual $\rightarrow$ for a new conditional connective to be discussed below.

6 A proof system for circular definitions

If a new conditional is needed to make sense of the logic in rule form that naturally follows the revision-theoretic account of circular definitions, then of course it is not enough to provide a new logical symbol to accomplish the task. We are indeed required to provide it with the proper logic. The goal of this section is precisely that: to devise a proof system whose axioms encapsulate the properties of the step conditionals and show that they are adequate to the semantics based on the revision–theoretic machinery. In doing so, I am relying on pre-existing work that can be used to fill the missing details, as well as clarifying the proof strategy[14].

The semantics that legitimates the two rules $(P \text{ in})_i$ and $(P \text{ out})_i$ above, is embodied at the formal level in systems $(\mathsf{C}_n)_{n\in\mathbb{N}}$ from [15, ch. 5], whose theorems correspond to formulas which are valid in the above sense of the expression[15]. These are systems of rules based on a language which uses labels for formulas to express validity at certain stages in a revision sequence: therefore, it contains expressions of the form θ^i to be read as "θ is valid at stage i (of a revision sequence)". By also exploiting this feature, I have introduced a new conditional $\rightarrow$ in [4] to turn systems $(\mathsf{C}_n)_{n\in\mathbb{N}}$ into axiom systems. The idea of that result now looks pretty straightforward since, if the language is equipped with those labels, then one only needs to add a dedicated conditional to express the passage from $\varphi(\overline{a}, P)^i$ to $P(\overline{a})^{i+1}$ from $(P \text{ in})_i$, and the converse passage from $(P \text{ out})_i$. In view of the said adequacy result, these two passages correspond to derivations within the appropriately chosen system from the family $(\mathsf{C}_n)_{n\in\mathbb{N}}$ of calculi. Therefore, it is enough for $\rightarrow$ to satisfy all those properties that are required to internalize them, as conditionals normally do (which explains why $\rightarrow$ satisfies the same logical principles as $\supset$, except for the fact that it applies to labelled formulas – see [4, §2]).

The available results makes the strategy to achieve the additional step I aim at here pretty simple too. Like $\rightarrow$, the step conditionals can be used to make sense of rules $(P \text{ in})_i$ and $(P \text{ out})_i$ by "comprising" those rules in a formula. Namely, assuming to maintain here Gupta and Standefer [16] notation and use $\rightarrow_1$ for the "step-down conditional" and $\leftarrow_1$ for the "step-up conditional", then

[14] For a much more comprehensive investigation related to the goal of this paper and the specific content of this section, I refer the reader to [5].

[15] In more precise terms, the idea of the extended validity relation explained in footnote 11 gives rise to a proper semantics which one can view as stratified into layers. The notion of regular hypotheses from §4 is made relative to their period leading, for every $n \in \mathbb{N}$, to the notion of *n-regular* hypothesis in a model $\mathbf{M}$ of the ground language (i.e., $\mathcal{L}^{-P} := \mathcal{L} \setminus \{P\}$) that applies to any hypothesis H_i (which is a subset of the domain of $\mathbf{M}$) for which $H_i = H_{i+n}$ is the case. This is used, in turn, to define the notion of *n-validity* for formulas of the full language $\mathcal{L}$, as the property of any θ for which there exists $m \in \mathbb{N}$ such that $(\mathbf{M}, H_{i+m}) \models \theta$ holds for every n-regular hypothesis H_i. The family of calculi $(\mathsf{C}_n)_{n\in\mathbb{N}}$ from [15, ch. 5] is such that, for each $n \in \mathbb{N}$, C_n is sound and complete with respect to this notion of n-validity (hence, every theorem of C_n is n-valid over any ground model $\mathbf{M}$, and every n-valid formula is provable in C_n). The special case for $n = 0$ of the n-validity relation applies to any formula of $\mathcal{L}$ for which there exists $m \in \mathbb{N}$ such that $(\mathbf{M}, H_{i+m}) \models \theta$ holds for *every* hypothesis H_i (since $H_i = H_{i+0}$ holds by definition).

by the formula $P(\overline{a}) \leftarrow_1 \varphi(\overline{a}, P)$ one expresses the content of $(P \text{ in})_i$, and by $P(\overline{a}) \rightarrow_1 \varphi(\overline{a}, P)$ the content of $(P \text{ out})_i$ instead. This is more than what one can do by making use of $\rightarrow$, since one needs labels for formulas to achieve the same in this case. That is, conditionals $\rightarrow_1$ and $\leftarrow_1$ embody the expressive power of $\rightarrow$ and the labels together. This simple observation is the key idea for the result that I present here, and which I now turn to develop in more precise terms.

Let, as in [4], $\mathcal{L}$ be any first-order language, $\mathcal{L}^+ := \mathcal{L} \cup \{P\}$ where $P(x)$ is a fresh unary predicate symbol, and $\mathcal{L}(I)$ being obtained from $\mathcal{L}^+$ by adding to it (i) the elements of the index-term set I, which contains terms $\overline{p}$ for every $p \in \mathbb{Z}$, and (ii) the said implication $\rightarrow$.[16] Labelled formulas were present already in the language of systems $(\mathsf{C}_n)_{n\in\mathbb{N}}$. As a matter of fact, it is safe to assume that these systems be also formulated in this language. Let instead $\mathcal{L}^{sc}$ be the language that is obtained from $\mathcal{L}^+$ by adding the step-down and the step-up conditionals $\rightarrow_1, \leftarrow_1$. Individual terms of both $\mathcal{L}(I)$ and $\mathcal{L}^{sc}$ are the same as terms of $\mathcal{L}^+$, while the set $FORM_I^+$ of formulas of $\mathcal{L}(I)$ is obtained by labelling the elements of the set $FORM^+$ of formulas of $\mathcal{L}^+$ with index terms from I (hence, $\phi^{\overline{p}}$ for $p \in \mathbb{Z}$ is a formula of $\mathcal{L}(I)$ whenever ϕ is a formula of $\mathcal{L}^+$), and then close the collection that is obtained in this way under possibly repeated applications of $\rightarrow$ (therefore, $\phi^{\overline{p}} \rightarrow \theta^{\overline{q}}$ is also a formula of $\mathcal{L}(I)$ whenever $\phi^{\overline{p}}, \theta^{\overline{q}}$ are). As far as the set $FORM_{sc}^+$ of formulas of $\mathcal{L}^{sc}$ is concerned instead, this is obtained by recursively closing the set $FORM^+$ under applications of conditionals $\rightarrow_1$ and $\leftarrow_1$[17].

The language $\mathcal{L}(I)$ was used in [4] to introduce a family $(\mathsf{HC}_n)_{n\in\mathbb{N}}$ of Hilbert-style axiom systems embodying the basic properties of finite revision sequences. Beside the logical properties of $\rightarrow$, which correspond to the axioms and rules for material implication of classical logic in labelled form (see [4, p. p. 919]), the axioms of these systems describe the validity properties for formulas which follow from the revision–theoretic semantics described above. So, for instance, the system HC_0 contains special axioms:

$$\begin{array}{ll} (DEF) & P(t)^{\overline{p+1}} \leftrightarrow \varphi(x, P)^{\overline{p}} \\ (IS) & \phi^{\overline{p}} \rightarrow \phi^{\overline{q}} \end{array}$$

for every $\overline{p}, \overline{q} \in I$, and for every formula ϕ of $\mathcal{L}$, i.e., not containing occurrences of the circular predicate P (and where $\phi^{\overline{p}} \leftrightarrow \theta^{\overline{q}}$ is an abbreviation for

[16] I am presenting here $\mathcal{L}(I)$ as a single language, whereas there were actually a family of them, $(\mathcal{L}(I)_n)_{n\in\mathbb{Z}}$, in [4]. This was motivated then by seeing each language $\mathcal{L}(I)_n$ as providing the corresponding theory HC_n I speake of below with its own linguistic base. Looking back, this dependence of languages on the parameter $n \in \mathbb{Z}$ was not really necessary, and it can be dropped as I do here for the sake of readability. I would like to thank an anonymous referee for giving me the opportunity of re-thinking this matter while I was trying to make up for a related comment she made on a previous draft of the paper.

[17] The acquainted reader would notice here that this is the same as adding the step conditionals as an "external feature" of the language $\mathcal{L}^{sc}$ to formulas of $\mathcal{L}^+$. In particular, $\rightarrow_1$ and $\leftarrow_1$ do not occur in the definiens $\varphi(x, P)$ of P. This is coherent with what was assumed in [4] and enough to achieve the modest goal of this paper, but is less than what is needed for the broader aim pursued by Gupta and Standefer (see [16, p. 47] in particular).

$(\phi^{\overline{p}} \to \theta^{\overline{q}}) \wedge (\theta^{\overline{q}} \to \phi^{\overline{p}}))$. In addition to the previous axioms, systems HC_n with $n > 0$ feature a generalization of (IS) to formulas of the full language, call it $(IS)_n$, which is justified by the existence of a period for finite revision sequences mentioned above:

$$(IS)_n \quad \phi^{\overline{p}} \leftrightarrow \phi^{\overline{p+n}}$$

for every $\overline{p} \in I$ and ϕ formula of $\mathcal{L}^+$.

The main result that certifies that systems $(\mathsf{HC}_n)_{n\in\mathbb{N}}$ are well-behaved with respect to the semantics of finite revision is the following theorem (Theorem 2.2 from [4, p. 921]) which states that they preserve derivability in systems $(\mathsf{C}_n)_{n\in\mathbb{N}}$ (hence, they are complete with respect to that semantics):

Theorem 1 *For every finite set Γ of formulas of $\mathcal{L}(I)$, and for every formula $\phi^{\overline{p}}$ of $\mathcal{L}(I)$, if $\Gamma \vdash_{\mathsf{C}_n} \phi^{\overline{p}}$, then $\Gamma \vdash_{\mathsf{HC}_n} \phi^{\overline{p}}$.*

As far as the goal of this paper is concerned, it is required only to make use of the following more restricted lemma (Lemma 2.3 from [4, p. 921]):

Lemma 1 *For all formulas ϕ, θ of $\mathcal{L}^+$, $\overline{p}, \overline{q} \in I$, and for all $n \in \mathbb{N}$*

$$\phi^{\overline{p}} \vdash_{\mathsf{C}_n} \theta^{\overline{q}} \Leftrightarrow \vdash_{\mathsf{HC}_n} \phi^{\overline{p}} \to \theta^{\overline{q}}$$

As a matter of fact, the latter result makes clear that the new implication connective $\to$ was meant, as I said, to internalize the derivability relation of systems $(\mathsf{C}_n)_{n\in\mathbb{N}}$ within the corresponding element of $(\mathsf{HC}_n)_{n\in\mathbb{N}}$, and that the goal was accomplished eventually.

Now, as I suggested, the relationship between $\mathcal{L}(I)$ and $\mathcal{L}^{sc}$ can be made pretty simple by letting $\to_1, \leftarrow_1$ doing all the job that is done by $\to$ and by the labels in the former language. In the special case represented by formulas that are instances of axiom (DEF) above, the whole idea is to let formulas

$$P(t) \to_1 \varphi(x, P)$$
$$P(t) \leftarrow_1 \varphi(x, P)$$

correspond in $\mathcal{L}^{sc}$ to formulas

$$P(t)^{\overline{p+1}} \to \varphi(x, P)^{\overline{p}}$$
$$\varphi(x, P)^{\overline{p}} \to P(t)^{\overline{p+1}}$$

of $\mathcal{L}(I)$. The example gives the occasion to further stress the spirit of the proposal: $\to_1, \leftarrow_1$ are regarded as syntactical devices that allow to express in $\mathcal{L}^{sc}$ a portion of the conditional reasonining that $\to$ *and* the labels express in $\mathcal{L}(I)$. The portion in question is the one that refers to any given stage and the preceding one as far as $\to_1$ is concerned, any stage and the successive one, instead, in the case of $\leftarrow_1$. This observation can be extended in order to provide a translation between formulas of the two languages that preserves derivability in systems $(\mathsf{HC}_n)_{n\in\mathbb{N}}$ within a suitably defined calculus, which relies on a non-labelled syntax.

The most direct way to achieve the goal is to supply the language $\mathcal{L}^{sc}$ with means to express the missing "layers" of the said conditional reasoning. Let then $\mathcal{L}^{sc}_n$ the language obtained from $\mathcal{L}^{sc}$ by adding to the list of its symbols a whole set $\{(\rightarrow_n, \leftarrow_n) | n \in \mathbb{N}\}$ of pairs of n-step conditionals. The set of formulas of this new language is obtained from the set of formulas of $\mathcal{L}^{sc}$ in the obvious way.

Now, assuming the previous correspondence to be suitably extended, and made use for the sake of a clause in the definition of a translation map from formulas of the form $\phi^{\overline{p+1}} \rightarrow \theta^{\overline{p}}$ and $\phi^{\overline{p}} \rightarrow \theta^{\overline{p+1}}$ of $\mathcal{L}(I)$ into formulas of $\mathcal{L}^{sc}_n$ of the form $\phi \rightarrow_1 \theta$ and $\phi \leftarrow_1 \theta$ respectively, an extended translation of the $\rightarrow$-fragment of $\mathcal{L}(I)$ into formulas of $\mathcal{L}^{sc}_n$ can be obtained by making use of the following additional correspondences:

$$\begin{array}{lcl} \phi^{\overline{p}} \rightarrow \theta^{\overline{q}} & \mapsto & \phi \rightarrow_{(p-q)} \theta, \text{ if } q < p \\ \phi^{\overline{p}} \rightarrow \theta^{\overline{q}} & \mapsto & \phi \leftarrow_{(q-p)} \theta, \text{ if } p < q \\ \phi^{\overline{p}} \rightarrow \theta^{\overline{p}} & \mapsto & \phi \supset \theta \end{array}$$

for every $p, q \in \mathbb{Z}$. These clauses are enough to extend the said translation to formulas of the form $\phi^{\overline{p}} \rightarrow \theta^{\overline{q}}$ where ϕ and θ are formulas of $\mathcal{L}^{+}$ (i.e., do not feature any further occurence of $\rightarrow$) and $\overline{p}, \overline{q}$ are labels whatsoever. For any such formula $\phi^{\overline{p}} \rightarrow \theta^{\overline{q}}$ of $\mathcal{L}(I)$, let $(\phi^{\overline{p}} \rightarrow \theta^{\overline{q}})^{\bullet}$ represent the corresponding formula of $\mathcal{L}^{sc}_n$. An inspection of systems $(\mathsf{HC}_n)_{n\in\mathbb{N}}$ quickly reveals that this extended translation is not enough to think of embedding these systems into axiomatic theories based upon language $\mathcal{L}^{sc}_n$, as among the logical axioms at use in $(\mathsf{HC}_n)_{n\in\mathbb{N}}$ there are formulas featuring nested occurrences of $\rightarrow$. The said translation can be extended to formulas of this sort though, by adapting the above idea accordingly as follows.

Let $\mathbf{p}, \mathbf{q}, \ldots$ stay for tuples of index terms, so that $\mathbf{p} = (\overline{p}_1, \ldots, \overline{p}_k), \mathbf{q} = (\overline{q}_1, \ldots, \overline{q}_h), \ldots$ and $p_i, q_j \in \mathbb{Z}$ for every $1 \leq i \leq k, 1 \leq j \leq h$. For the sake of readability, let us also make use of some standard notation for tuples, and indicate with $(\mathbf{p})_i$ the i-th element in the tuple $\mathbf{p}$ (i.e., $\mathbf{p} = ((\mathbf{p})_0, \ldots, (\mathbf{p})_k)$). Tuples of index terms were used in [4] to define the full $\rightarrow$-fragment of $\mathcal{L}(I)$, whose elements may contain nested occurrences of the implication connective $\rightarrow$ up to different depths in both the antecedent and the consequent. Therefore, the main subformulas of a formula featuring $\rightarrow$ as principal logical operation are allowed to carry different labels. I used there the notation $\phi^{\mathbf{p}} \rightarrow \theta^{\mathbf{q}}$ to indicate the fact that ϕ and θ feature possibily nested occurrences of the symbol $\rightarrow$, and $\mathbf{p}, \mathbf{q}$ collect together labels of subformulas to which $\rightarrow$ applies in the order these formulas appear in the construction tree of ϕ and θ themselves.

Granted that, let the $\rightarrow$-*degree* of formulas of the $\rightarrow$-fragment of $\mathcal{L}(I)$ be inductively defined as follows:

- formulas of $\mathcal{L}(I)$ with $\rightarrow$-degree 0 are all and only formulas of the form $\phi^{\overline{p}} \rightarrow \theta^{\overline{q}}$ where ϕ, θ are formulas of $\mathcal{L}^{+}$ whatsoever, and $\overline{p}, \overline{q} \in I$;
- formulas of $\mathcal{L}(I)$ with $\rightarrow$-degree $n+1$ are all formulas of degree $k \leq n$ plus formulas of the form $\phi^{\mathbf{p}} \rightarrow \theta^{\mathbf{q}}$ where $\phi^{\mathbf{p}}, \theta^{\mathbf{q}}$ are any formulas of $\mathcal{L}(I)$ with $\rightarrow$-degree n, and $(\mathbf{p})_i, (\mathbf{q})_i \in I$ for every $0 \leq i \leq n$.

That is: formulas with $\rightarrow$-degree 0 feature no nested occurences of $\rightarrow$ in their subformulas; formulas with $\rightarrow$-degree 1 feature occurrences of $\rightarrow$ in their immediate subformulas but not in any subformula of them, and so on. In particular, formulas of $\mathcal{L}(I)$ with degree 2 are formulas of the form $\phi^{\mathbf{p}} \rightarrow \theta^{\mathbf{q}}$ where $\phi^{\mathbf{p}}, \theta^{\mathbf{q}}$ are either of degree 0 and feature no further occurrence of $\rightarrow$, or they are of degree 1, hence they are of the form $\eta^{\mathbf{r}} \rightarrow \rho^{\mathbf{s}}$ where $\eta^{\mathbf{r}}, \rho^{\mathbf{s}}$ are again of degree 0, or are of degree 1 and feature occurrences of $\rightarrow$ in their immediate subformulas, but no more than that. Formulas of this portion of the $\rightarrow$-fragment of $\mathcal{L}(I)$ are precisely those to which we need to extend the above translation for the modest purpose here at stake[18]. By the way, this goal can be achieved by considering these further correspondences[19]:

$$\begin{array}{lcl}
(\phi^{\overline{p}} \rightarrow \theta^{\overline{q}}) \rightarrow \eta^{\overline{r}} & \mapsto & (\phi^{\overline{p}} \rightarrow \theta^{\overline{q}})^{\bullet} \rightarrow_{(|p-q|-r)} \eta, \text{ if } r < |p-r| \\
(\phi^{\overline{p}} \rightarrow \theta^{\overline{q}}) \rightarrow \eta^{\overline{r}} & \mapsto & (\phi^{\overline{p}} \rightarrow \theta^{\overline{q}})^{\bullet} \leftarrow_{(r-|p-q|)} \eta, \text{ if } r > |p-r| \\
(\phi^{\overline{p}} \rightarrow \theta^{\overline{q}}) \rightarrow (\eta^{\overline{r}} \rightarrow \nu^{\overline{s}}) & \mapsto & (\phi^{\overline{p}} \rightarrow \theta^{\overline{q}})^{\bullet} \rightarrow_{(|r-s|-|p-q|)} (\eta^{\overline{r}} \rightarrow \nu^{\overline{s}})^{\bullet}, \text{ if } |r-s| > |p-r| \\
(\phi^{\overline{p}} \rightarrow \theta^{\overline{q}}) \rightarrow (\eta^{\overline{r}} \rightarrow \nu^{\overline{s}}) & \mapsto & (\phi^{\overline{p}} \rightarrow \theta^{\overline{q}})^{\bullet} \leftarrow_{(|p-q|-|r-s|)} (\eta^{\overline{r}} \rightarrow \nu^{\overline{s}})^{\bullet}, \text{ if } |r-s| < |p-r|
\end{array}$$

and by addressing in the obvious manner the symmetrical cases.

Let us now indicate by symbols $\phi^{\mathbf{p}}, \theta^{\mathbf{q}}, \ldots$ formulas whatsoever of the $\rightarrow$-fragment of $\mathcal{L}(I)$ with $\rightarrow$-degree 2. Let, for any formula $\phi^{\mathbf{p}}$ of that sort, $(\phi^{\mathbf{p}})^{\bullet}$ be the formula of $\mathcal{L}_n^{sc}$ that corresponds to it according to the clauses listed above. In turn, this can be used to introduce unlabelled version uHC_n of systems $(\mathsf{HC}_n)_{n\in\mathbb{N}}$ as follows:

- $(\phi^{\mathbf{p}})^{\bullet}$ is an axiom of uHC_n for every $\varphi^{\mathbf{p}}$ which is an axiom of any of systems $(\mathsf{HC}_n)_{n\in\mathbb{N}}$;
- uHC_n features the following rules of inference

$$\frac{\phi \rightarrow_k \theta(x),\ x \notin FV(\phi)}{\phi \rightarrow_k \forall x\theta} \qquad \frac{\theta(x) \rightarrow_k \phi,\ x \notin FV(\phi)}{\exists x\theta \rightarrow_k \phi} \qquad \frac{\phi \rightarrow_k \theta \quad \phi}{\theta}$$

$$\frac{\phi \leftarrow_k \theta(x),\ x \notin FV(\phi)}{\phi \leftarrow_k \forall x\theta} \qquad \frac{\theta(x) \leftarrow_k \phi,\ x \notin FV(\phi)}{\exists x\theta \leftarrow_k \phi} \qquad \frac{\phi \leftarrow_k \theta \quad \phi}{\theta}$$

  for every $k \in \mathbb{N}$ and for every formula ϕ, θ of $\mathcal{L}_n^{sc}$.

Then, the following has an easy proof:

Proposition 1 *Let $\phi^{\mathbf{p}}$ be any formula of the $\rightarrow$-fragment of $\mathcal{L}(I)$ with $\rightarrow$-degree 2. Then, for every and $n \in \mathbb{N}$*

$$\vdash_{\mathsf{HC}_n} \phi^{\mathbf{p}} \Leftrightarrow \vdash_{\mathsf{uHC}_n} (\phi^{\mathbf{p}})^{\bullet}$$

Proof By induction on the length of the given derivation. ■

[18] An inspection of results and proofs from [4] shows that formulas with $\rightarrow$-degree equal to 3 or more, get involved only in results which are independent of Lemma 1 above (in particular, they are required to prove the deduction theorem for $\rightarrow$ – see [4, Thrm. 2.1]).

[19] For $p, q \in \mathbb{Z}$, I use the standard notation $|p-q|$ to use the modulus operation over integer numbers (hence, $|p-q| = (p-q)$ if $p \geq q$, $|q-p| = (q-p)$ if $q > p$).

In view of the said adequacy of systems $(\mathsf{HC}_n)_{n\in\mathbb{N}}$ with respect to the semantics of finite revision, owing to Lemma 1 above and this latter result it can be argued that this feature is ineherited by the unlabelled systems $(\mathsf{uHC}_n)_{n\in\mathbb{N}}$.

7 Conclusion

In this paper I have discussed the issue of predicates defined circularly, which I have tried to motivate by presenting it as a chapter in the traditional account of definitions that has only recently found its space in philosophical debates. As I have argued for here, this lack of prior consideration of this topic seems to me to be due to two main factors: (i) the received view of definitions from classical philosophy that simply prevented the circular case to be even considered as a theoretical possibility, up to recent times; (ii) the lack of actual cases of circular definitions that might motivate a shift from that view. From the point of view of the actual practice, this latter aspect can be explained as an oversight as the examples I discussed in §3.1 are commonly at use in everyday life and could have engendered some reflections on the topic even before some other reasons prompted the case of circular definitions to be raised and analyzed. As a matter of fact, the main motivation that let this issue gradually gain attention among scholars comes from the deepening of the philosophical investigation by formal methods over topics like truth in formal languages, and rational and strategic choice in finite games. This treatment of abstract concepts coheres with the view that I have exemplified here by Bolzano's quote from §2. By referring to recent work in the area, I have tried to argue in §§4-5 that an analysis of circular definitions along this line of thoughts naturally leads to the development of proof systems that may encapsulate the logic that underlies them, in order to compare it with what happens in the standard, non-circular case. Finally, in §6 I have sketched a proposal in this respect, by presenting the system uHC_n that turns out to be adequate to the natural semantics of circular definitions on the one hand, and which features new conditionals from Gupta and Standefer [16] that are bound to express the conditional forms of reasoning associated to that semantics.

Some further issues of a more technical character naturally stem from this presentation, like, for example: whether the relationship between the 1-step conditionals and the n-steps ones I have considered here can be further clarified, and possibly lead to a definition of conditionals of the latter forms in terms of those of the former ones; or, whether a proof-theoretically palatable version of system uHC_n can also be found (as it was the case for Gentzen-style versions $(\mathsf{GC}_n)_{n\in\mathbb{N}}$ of systems $(\mathsf{HC}_n)_{n\in\mathbb{N}}$ in [4]), and similar other problems. Some of them are matter of my current research concerns, and new results about them will be eventually published elsewhere.

References

[1] Conrad Asmus. Vagueness and revision sequences. *Synthese*, 190:953–974, 2013.

[2] Bernard Bolzano. Contribution to a better-grounded presentation of mathematics. In W. Ewald, editor, *From Kant to Hilbert: A Source Book in the Foundations of Mathematics*, pages 174–224. Clarendon Press, Oxford, 1996.

[3] Lesley Brown. Definition and division in Plato's *sophist*. In D. Charles, editor, *Definitions in Greek Philosophy*, pages 151–171. Oxford University Press, 2010.

[4] Riccardo Bruni. Analytic calculi for circular concepts by finite revision. *Studia Logica*, 101:915–932, 2013.

[5] Riccardo Bruni. Proof-theoretic analysis of step conditionals. In preparation, 2017.

[6] Davis Charles. Definition and explanation in the *posterior analytics* and *metaphysics*. In D. Charles, editor, *Definitions in Greek Philosophy*, pages 286–328. Oxford University Press, 2010.

[7] Davis Charles. The paradox in the *meno* and Aristotle's attempts to resolve it. In D. Charles, editor, *Definitions in Greek Philosophy*, pages 115–150. Oxford University Press, 2010.

[8] Kei Chiba. Aristotle on essence and defining-phrase in his dialectic. In D. Charles, editor, *Definitions in Greek Philosophy*, pages 172–199. Oxford University Press, 2010.

[9] David Charles (editor). *Definitions in Greek Philosophy*. Oxford University Press, 2010.

[10] Solomon Feferman. The proof theory of classical and constructive inductive definitions. a 40 year saga, 1968-2008. In R. Schindler, editor, *Ways of Proof Theory*, pages 7–30. Ontos Verlag, 2010.

[11] Mary Louise Gill. Division and definition in Plato's *sophist* and *statestman*. In D. Charles, editor, *Definitions in Greek Philosophy*, pages 172–199. Oxford University Press, 2010.

[12] Anil Gupta. Definition and revision: A response to McGee and Martin. *Philosophical Issues*, 8:419–443, 1997.

[13] Anil Gupta. On circular concepts. In *Truth, Meaning and Experience*, pages 95–134. Oxford University Press, 2011.

[14] Anil Gupta. Definitions. *Stanford Encyclopedia of Philosophy*, `https://plato.stanford.edu/entries/definitions`, revised on April 2015.

[15] Anil Gupta and Nuel Belnap. *The Revision Theory of Truth*. MIT Press, 1993.

[16] Anil Gupta and Shawn Standefer. Conditionals in theories of truth. *Journal of Philosophical Logic*, 46:27–63, 2017.

[17] Donald A. Martin. Revision and its rivals. *Philosophical Issues*, 8:407–418, 1997.

[18] Vann McGee. Revision. *Philosophical Issues*, 8:387–406, 1997.

[19] Alfred Tarski. The concept of truth in formalized languages. In J. Corcoran, editor, *Logic, semantics and metamathematics*, pages 152–278. Hackett, Indianapolis, 1983 (2nd edition).